DOCUMENT RESUME

ED 134 926 CS 003 178

AUTHOR Halff, Henry M.
TITLE Graphical Evaluation of Hierarchical Clustering Schemes. Technical Report No. 1.
INSTITUTION Illinois Univ., Urbana. Center for the Study of Reading.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Oct 75
GRANT NIE-G-74-0007
NOTE 11p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Classification; *Cluster Analysis; Cluster Grouping; *Graphs; *Models; *Statistical Analysis
IDENTIFIERS *Center for the Study of Reading (Illinois)

ABSTRACT

Graphical methods for evaluating the fit of Johnson's hierarchical clustering schemes are presented together with an example. These evaluation methods examine the extent to which the clustering algorithm can minimize the overlap of the distributions of intracluster and intercluster distances. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED134926

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Technical Report No. 1

GRAPHICAL EVALUATION OF HIERARCHICAL CLUSTERING SCHEMES

Henry M. Halff

October 1975

UNIVERSITY OF ILLINOIS at Urbana-Champaign
LABORATORY FOR COGNITIVE STUDIES IN EDUCATION
236 Education Building
Urbana, Illinois 61801

CS 003 178

The research reported herein was supported in part by the National Institute of Education under Contract HEW NIE-G-74-0007. The author would like to thank Halff Associates of Dallas, Texas for preparing the figure.

Abstract

Graphical methods for evaluating the fit of Johnson's hierarchical clustering schemes are presented together with an example. These evaluation methods examine the extent to which the clustering algorithm can minimize the overlap of the distributions of intracluster and intercluster distances.

Graphical Evaluation of Hierarchical Clustering Schemes

A general problem with cluster analysis is that of evaluating the appropriateness of the method for any particular set of data. This paper presents such a method for Johnson's [1967] hierarchical clustering schemes.

Johnson suggests a model for cluster analysis in which the terms to be clustered are the endpoints of a tree. Each node or intersection of branches is assigned a numerical value or level which increases monotonically with the node's distance from the endpoints. If we let the distance between any two points be the level of the node first joining those points, then a necessary and sufficient condition for the existence of the specified tree is that the interpoint distances be ultrametric; that is,

$$(1) \qquad \underline{d}(\underline{x},\underline{z}) \leq \max[\underline{d}(\underline{x},\underline{y}), \underline{d}(\underline{y},\underline{z})],$$

where $\underline{d}(\underline{x},\underline{y})$ is the distance between the points $\underline{x}$ and $\underline{y}$.

The construction of a cluster tree from a set of ultrametric distances is simple, but, in many cases, we cannot expect the distances to conform exactly to (1). Johnson suggests two methods for cluster analysis when $\underline{d}$ is not ultrametric. Both methods involve starting with each term in a separate cluster and then using an iterative procedure to join the clusters until only a single cluster remains. Each iteration involves first joining those clusters which have the lowest intercluster distance and then determining the distance of the newly formed cluster from the other clusters. It is this second step of each iteration that differentiates the two methods suggested by Johnson. Let $\underline{D}$ be the intercluster

distance function, and suppose that a new cluster, $\underline{x}$, is formed on some iteration by joining the clusters $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$. Then for the maximum method

$$D(\underline{x}, \underline{y}) = \max[\underline{D}(\underline{x}_1,\underline{y}), \underline{D}(\underline{x}_2,\underline{y}), \ldots, \underline{D}(\underline{x}_n,\underline{y})].$$

For the minimum method

$$\underline{D}(\underline{x},\underline{y}) = \min[\underline{D}(\underline{x}_1,\underline{y}), \underline{D}(\underline{x}_2,\underline{y}), \ldots, \underline{D}(\underline{x}_n,\underline{y})].$$

In either case, the construction of the tree is simple. Each node represents the formation of a new cluster from lower level clusters, and the level of that node is simply the intercluster distance of the clusters making up the new cluster.

Our suggestion for evaluating each method is a goodness-of-fit index which may be formulated for each level of the tree. To introduce this method, first consider the sets of clustered and unclustered points at each level, $\underline{v}$. Letting $\delta(\underline{x},\underline{y})$ be the lowest level at which $\underline{x}$ and $\underline{y}$ cluster, $\underline{C}_v = \{(\underline{x},\underline{y});\ \delta(\underline{x},\underline{y}) \leq \underline{v})\}$ is the set of all pairs of points clustered at or below Level $\underline{v}$. The complementary set, those clustered above Level $\underline{v}$, is $\underline{U}_v = \{(\underline{x},\underline{y});\ \delta(\underline{x},\underline{y}) > \underline{v}\}$. It is not difficult to see that for either method $\underline{d}(\underline{x},\underline{y}) \leq \underline{v}$ iff $(\underline{x},\underline{y}) \in \underline{C}_v$ whenever (1) holds. But when (1) does not hold, the two methods make different compromises with this condition. Specifically, the maximum method insures that $\underline{d}(\underline{x},\underline{y}) \leq \underline{v}$ for any $(\underline{x},\underline{y}) \in \underline{C}_v$. Conversely, the minimum method insures that $\underline{d}(\underline{x},\underline{y}) > \underline{v}$ for any $(\underline{x},\underline{y}) \in \underline{U}_v$. Departures of $\underline{d}$ from (1) can therefore be evaluated for the maximum method by computing the proportion, $\underline{P}_U(\underline{v})$,

of (x,y) pairs in U_v for which $d(x,y) \leq v$. The corresponding statistic for the minimum method is the proportion, $P_C(v)$, of (x,y) pairs in C_v for which $d(x,y) > v$. Both of these statistics, $P_U(v)$ and $P_C(v)$ will exceed zero to the extent that (1) is violated in the data.

As an example, Figure 1 presents $P_U(v)$ and $P_C(v)$ as a function of v for Johnson's analysis of the data presented in Table VII of Miller and Nicely [1955]. This table gives confusion counts among 16 consonants, and from these counts Johnson constructed a similarity function, $s(x,y)$, which is presumably related to $d(x,y)$ by a decreasing monotonic function. Thus, $P_U(v)$ in this case is the proportion of (x,y) pairs in U_v for which $s(x,y) \geq v$, and $P_C(v)$ is the proportion of (x,y) pairs in C_v for which $s(x,y) < v$. Plotted with $P_U(v)$ is an obvious standard of bad fit, the unconditional proportion, $F(v)$ of (x,y) pairs for which $s(x,y) \geq v$; $P_U(v)$ would be expected to approach $F(v)$ only under the worst of conditions. The corresponding standard of bad fit for the minimum method is $1 - F(v)$, which is shown for this example with the plot of $P_C(v)$.

Insert Figure 1 about here

Apart from a goodness-of-fit interpretation, P_U can be viewed as an index of the overall difference of within- and between-cluster distances in that $P_U(v)$ is the proportion of intercluster distances not exceeding the highest intracluster distance. The converse unfortunately does not hold for $P_C(v)$ since this statistic can exceed the proportion of intracluster distances greater than or equal to the lowest intercluster distance. This latter proportion is plotted as a function of

level in Figure 1, where it is designated $\underline{P}'_C(\underline{v})$. Applying this interpretation to Johnson's example, we see that the maximum method is quite successful at limiting large distances to intercluster pairs. The minimum method is less consistent in limiting small distances to intercluster pairs. Furthermore, the last two nodes formed by both methods appear to be more or less meaningless.

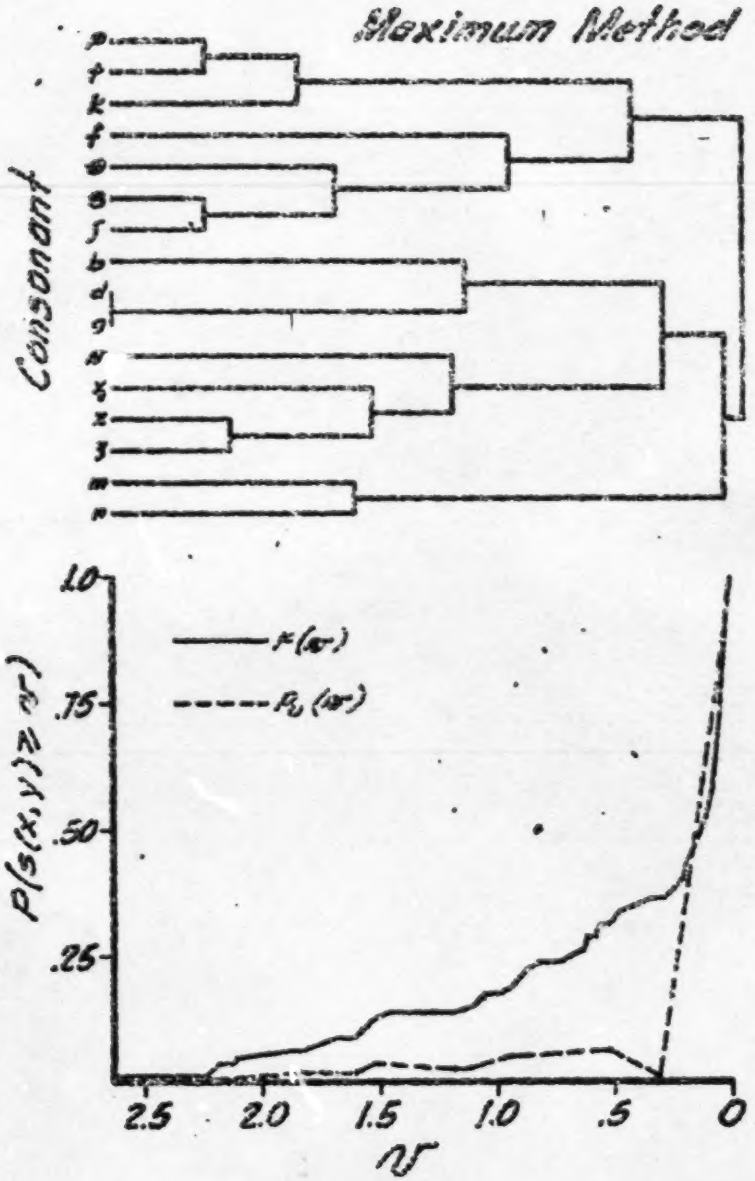
References

Johnson, S. C. Hierarchical clustering schemes. Psychometrika, 1967, 32, 241-254.

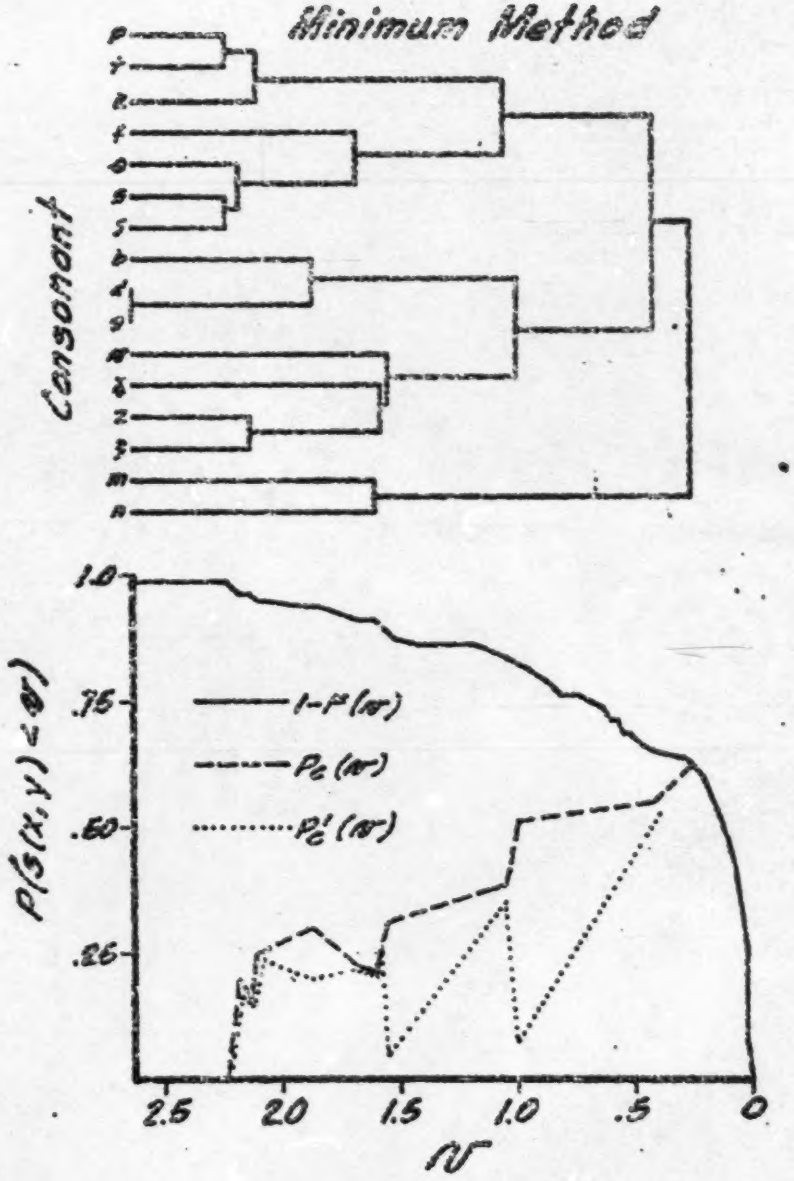
Miller, G. A., & Nicely, P. E. An analysis of perceptual confusions among some English consonants. Journal of the Acoustical Society of America, 1955, 27, 338-352.

Figure Caption

Figure 1. Cluster trees and evaluation functions for Johnson's analysis of Miller and Nicely's Table VII.


Maximum Method
Consonant
p
t
k
f
θ
s
ʃ
b
d
g
v
ð
z
ʒ
m
n
1.0
.75
.50
.25
P(S(x,y)≥ r)
F(r)
Pc(r)
2.5
2.0
1.5
1.0
.5
0
r

10


Minimum Method
Consonant
p
t
k
f
θ
s
ʃ
b
d
g
v
ð
z
ʒ
m
n
1.0
.75
.50
.25
P(S(x,y)< r)
1-F(r)
Pc(r)
Pc'(r)
2.5
2.0
1.5
1.0
.5
0
r

11